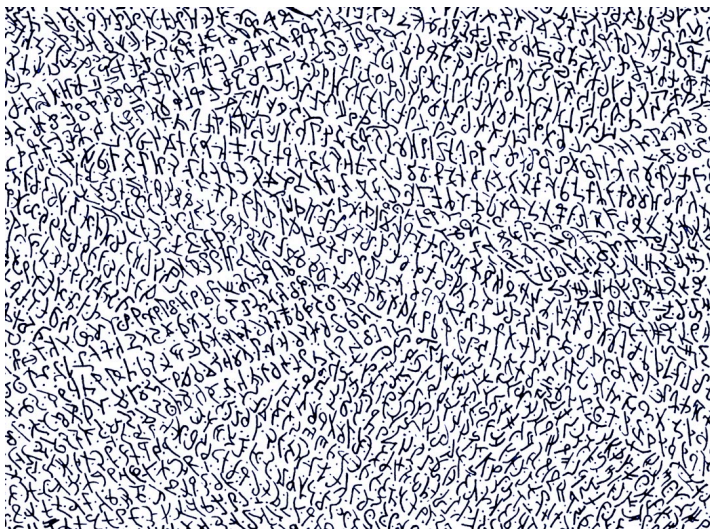


# Corpus et constitution des savoirs linguistiques

## RÉSUMÉS DES COMMUNICATIONS



E. Demont-Suisse [apozepoa@gmail.com](mailto:apozepoa@gmail.com)



30 janvier 2015 :  
INALCO/BULAC, amphithéâtre : 65, rue des Grands Moulins, 75013 Paris  
31 janvier 2015 :  
BNF, salle 70: Quai François Mauriac, 75013 Paris

## DONNEES AUTHENTIQUES : UN GRAND CORPUS DE SMS EN FRANÇAIS

Rachel PANCKHURST\*, Mathieu ROCHE\*\*, Cédric LOPEZ\*\*\*

\* Praxiling UMR 5267 CNRS & Université Paul-Valéry Montpellier 3

[rachel.panckhurst@univ-montp3.fr](mailto:rachel.panckhurst@univ-montp3.fr)

\*\* Tetis, MTD (Maison de la Télédétection) UMR TETIS

[mathieu.roche@cirad.fr](mailto:mathieu.roche@cirad.fr)

\*\*\* Viseo Technologies

[cedric.lopez@viseo.com](mailto:cedric.lopez@viseo.com)

*À Augusta Mela,  
en mémoire de son œuvre interdisciplinaire entre linguistique et informatique*

Qu'est-ce que la donnée écrite en sciences du langage ? Trois types se distinguent : 1) la *donnée lexicale*, qui se présente essentiellement sous forme d'une entrée lexicale, regroupant un ensemble de propriétés ; 2) « le nom spécifique de la donnée observable en linguistique est l'*exemple* » et renvoie à « un énoncé qui pourrait être effectivement prononcé, même s'il ne l'est pas dans les faits » (Milner 1989, p. 51-52) ; 3) la donnée en tant que texte brut, i.e. *le corpus*. En linguistique(s) de corpus, il s'agit d'analyser les productions *authentiques* contenues dans le corpus. Dans certaines écoles linguistiques, au contraire, l'étude du corpus tout-venant n'a pas lieu d'être. Ainsi, perdure le débat concernant l'opposition (ou, tout au moins, la différenciation) entre exemples linguistiques (éventuellement « fabriqués ») et productions authentiques relevées dans des corpus (*cf.* entre autres, pour le français, Bilger *et al.* 2000, Cori *et al.* 2008, Habert *et al.* 1997, Péry-Woodley 1995).

En vingt ans, notre propre approche a évolué : d'une analyse linguistique-informatique basée sur l'*exemple* (Panckhurst 1994, p. 39), nous sommes passée à une analyse de la donnée *authentique* figurant dans des corpus (Panckhurst 2013, p. 97, Panckhurst *et al.* 2014). Pour nous, cette mutation s'explique, d'une part, par l'évolution de l'accès aux données, et, d'autre part, par le *discours électronique médié* (Panckhurst 1997, 2006), circulant entre individus se servant d'outils électroniques (ordinateurs, tablettes, téléphones portables, etc.), qui induit des pratiques et des usages émergents. En deux décennies, la constitution de corpus numérisés ou nativement numériques est devenue monnaie courante, et cette accessibilité massive constitue en soi une nouveauté. Les données authentiques existant sous la forme de courriels, forums, chats, blogs, réseaux sociaux, et, plus récemment de SMS, facilement exploitables par les chercheurs, permettent l'observation, la fouille et l'analyse des pratiques et des usages (novateurs ou non) des scripteurs.

Dans le cadre de cette communication, nous expliquerons ce cheminement, en nous focalisant sur des recherches récentes, portant sur le recueil, le traitement et l'analyse d'un grand corpus de SMS en français, intitulé « 88milSMS » (consultable sur la grille de services d'Huma-Num). En 2004, des universitaires belges ont lancé un projet international, *sms4science* ([www.sms4science.org](http://www.sms4science.org), Fairon *et al.* 2006, Cougnon 2015), afin de constituer une grande base de données mondiale de SMS authentiques. D'autres collectes ont suivi : en 2011, plus de 93 000 SMS ont été recueillis auprès du grand public (qui pouvait également répondre à un questionnaire sociolinguistique) par un groupe de chercheurs dans la région Languedoc-Roussillon (projet *sud4science LR*, [www.sud4science.org](http://www.sud4science.org), Panckhurst *et al.* 2013, Panckhurst & Moïse 2014). À l'aide d'exemples extraits de « 88milSMS », nous montrerons que les données peuvent être appréhendées selon deux approches, « fondée sur corpus » ('corpus-based') et « guidée par corpus » ('corpus-driven'), et que le va-et-vient constant entre les hypothèses et l'observation des données constitue le point essentiel de notre démarche. L'élaboration de ce corpus a participé au développement d'un logiciel

d'anonymisation semi-automatique, *Seek&Hide*, par des étudiants (Accorsi *et al.* 2014, Patel *et al.* 2013), et d'un prototype, permettant la construction automatique de dictionnaires électroniques de SMS selon une méthode d'alignement statistique (Lopez *et al.* 2014).

## Mots clefs

CORPUS, SMS, DONNEES AUTHENTIQUES, TAL, DISCOURS ELECTRONIQUE MEDIE, LOGICIEL D'ANONYMISATION, DICTIONNAIRES ELECTRONIQUES, ALIGNEMENT

## Éléments bibliographiques

- ACCORSI P., PATEL N., LOPEZ C., PANCKHURST R., ROCHE M. 2014. « Seek&Hide: Anonymising a French SMS corpus using natural language processing techniques », L.-A. Cougnon, C. Fairon ed., *SMS Communication. A Linguistic Approach*, Amsterdam, Philadelphia, John Benjamins, p. 11-28.
- BILGER M. 2000. *Corpus : Méthodologie et applications linguistiques*, Paris, Champion.
- CORI M., DAVID S., LEON J. éd. 2008. « Construction des faits en linguistique : la place des corpus », *Langages*, 171, Paris, Larousse, septembre 2008.
- COUGNON L.-A. à paraître, 2015. *Langage et sms. Une étude internationale des pratiques actuelles*. Presses universitaires de Louvain.
- FAIRON C., KLEIN J.-R., PAUMIER S. 2006. *SMS pour la science. Corpus de 30.000 SMS et logiciel de consultation*, Presses universitaires de Louvain, Louvain-la-Neuve, Manuel + CD-Rom, <http://www.smspourscience.be/>
- HABERT, B., NAZARENKO, A., & SALEM, A. 1997. *Les linguistiques de corpus*, Paris, Armand Colin.
- LOPEZ C., BESTANDJI R., ROCHE M., PANCKHURST R. 2014. "Towards Electronic SMS Dictionary Construction: An Alignment-based Approach", Actes du colloque LREC, Reykjavik, Islande, May 26-31, 2833-2838.  
[www.lrec-conf.org/proceedings/lrec2014/pdf/753\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2014/pdf/753_Paper.pdf)
- MILNER, J.-C. 1989. *Introduction à une science du langage*, Paris, Seuil.
- PANCKHURST R. 1994. "A Database for linguists: intelligent querying and increase of data", *Computers and the Humanities*, 28, p. 39-52.
- PANCKHURST R. 1997. « La communication médiatisée par ordinateur ou la communication médiée par ordinateur ? », *Terminologies nouvelles*, 17, p. 56-58.
- PANCKHURST R. 2006. « Le discours électronique médié : bilan et perspectives », A. Piolat éd. *Lire, écrire, communiquer et apprendre avec Internet*, Marseille, Éditions Solal, p. 345-366.
- PANCKHURST R. 2013. "A large SMS corpus in French : from design and collation to anonymisation, transcoding and analysis", Colloque CILC2013, Alicante, 14-16 mars : <http://web.ua.es/en/cilc2013/>, Actes du colloque, Procedia — Social and Behavioural Sciences, Elsevier.  
<http://www.sciencedirect.com/science/article/pii/S1877042813041475>
- PANCKHURST R. et MOÏSE C. 2014, "French text messages. From SMS data collection to preliminary analysis", L.-A. Cougnon, C. Fairon, ed., *SMS Communication. A Linguistic Approach*, Amsterdam/Philadelphia, John Benjamins, p. 141-168.
- PANCKHURST R., DETRIE C., LOPEZ C., MOÏSE C., ROCHE M., VERINE B. 2013. « Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS », *Épistémè — revue internationale de sciences sociales appliquées*, 9 : « Des usages numériques aux pratiques scripturales électroniques », p. 107-138.
- PANCKHURST R., DETRIE C., LOPEZ C., MOÏSE C., ROCHE M., VERINE B. 2014. "88milSMS. A corpus of authentic text messages in French", produit par l'Université Paul-Valéry Montpellier 3 et le CNRS, en collaboration avec l'Université catholique de Louvain,

financé grâce au soutien de la MSH-M et du Ministère de la Culture (Délégation générale à la langue française et aux langues de France) et avec la participation de Praxiling, Lirmm, Lidilem, Tetis, Viseo. ISLRN : 024-713-187-947-8.

- PATEL N., ACCORSI P., INKPEN D., LOPEZ C., ROCHE M. 2013. “Approaches of anonymisation of an SMS corpus”, Alexander Gelbukh ed., *Computational Linguistics and Intelligent Text Processing*, 14th International Conference, CICLing 2013, Samos, Greece, Berlin, Heidelberg, Springer Verlag, p. 77-88.
- PERY-WOODLEY, M.-P. 1995. « Quels corpus pour quels traitements automatiques ? », *T.A.L.*, 36/1-2, p. 213-232.